Department of Computer Science,
Faculty of Mathematical Sciences
University of Delhi
110007

MCSC105: DATA MINING
SEMESTER 1 PROJECT

AIR QUALITY PREDICTION

Submitted to:
Dr. Bharti Rana
Associate Professor
Department of Computer Science
Faculty of Mathematical Sciences
University of Delhi

Submitted by:
Prachi Bhatia
28
MSc Computer Science
Department of Computer Science
Faculty of Mathematical Sciences
University of Delhi

Introduction

Delhi, the capital city of India, spans an area of 1,483 km² and had a population of 16.9 million as of 2007-08. The city's rapid industrialization and migration trends have led to the presence of approximately 5.6 million vehicles, as reported in the Economic Survey of Delhi (2008–09). Delhi also boasts one of the highest road densities in India, with 1,749 km of road per 100 km². The combination of rapid population growth and economic development has placed immense strain on the city's transportation infrastructure, resulting in severe issues such as air pollution, traffic congestion, and reduced productivity. Additionally, the city grapples with substantial environmental challenges, including the accumulation of legacy waste at three major landfill sites—Bhalswa, Okhla, and Ghazipur. Compounding these issues are inadequate waste management practices, which exacerbate the environmental and public health concerns.

Air Quality Index

Air Quality Index is a tool for effective communication of air quality status to citizens in terms which are easy to understand. It transforms complex air quality data of various pollutants into an index value, nomenclature, and colour.

AQI calculation

There are six AQI categories namely Good,Satisfactory, Moderately Polluted, Poor, Very Poor, and Severe, Each of these categories is decided based on ambient concentration values of air pollutants and their likely health impacts (known as health breakpoints). AQ sub-index and health breakpoints are evolved for eight pollutants (PM10, PM2.5, NO2, SO2, CO, O3, NH3, and Pb) for which short term (upto 24-hours, 8-hourly in case of CO and O3) National Ambient Air Quality Standards are prescribed. The worst sub-index is the AQI for that location. The Individual pollutant-wise sub-index will provide air quality status for that pollutant.

Overall AQI is calculated only if data are available for minimum three pollutants out of which one should necessarily be either PM2.5 or PM10. Else, data are considered insufficient for calculating AQI. Similarly, a minimum of 16 hours' data is considered necessary for calculating subindex.

Project Objective

The project aims to achieve the following:

1. Inform Delhi public regarding overall status of air quality through a summation parameter that is easy to understand. It's useful for people who suffer from illness aggravated or caused by air pollution. Thus it enables them to modify their daily activities at times when they are informed of high pollution levels.
2. To assist in comparing air quality conditions at different locations within the National Capital. Thus, pointing out areas and frequencies of potential hazards.
3. To determine change in air quality (degradation or improvement) which have occurred over a specified period. This enables forecasting of air quality and plan pollution control measures.
4. Scientific Research: As a means for reducing a large set of data to a comprehendible form that gives better insight to the researcher while conducting a study of some environmental phenomena. This enables more objective determination of the contribution of individual pollutants and sources to overall air quality. Such tools become more useful when used in conjunction with other sources such as local emission surveys.

The scope of this project lies in identifying and ranking the underlying causes of severely poor historical and future (as forecasted) air quality standards of the Union Territory. Comprehension of such data will allow the public to press on issues requiring urgent dissemination of public policy; target fields will essentially include waste management, vehicular emissions, and industry.

Related Work

Air Quality Index (AQI) prediction is a critical aspect of urban environmental management, particularly in cities like Delhi, which face significant air pollution issues. Recent studies have explored various machine learning and statistical approaches to improve the accuracy of AQI forecasting. Among these, ARIMA (AutoRegressive Integrated Moving Average) and PCR (Principal Component Regression) have emerged as useful methods, both individually and in combination, for predicting AQI levels.

The prediction of AQI using a combination of ARIMA (AutoRegressive Integrated Moving Average) and PCR (Principal Component Regression) offers both statistical robustness and flexibility in handling complex datasets. ARIMA, a widely used time series forecasting method, operates on the principle of modeling past values of a series to predict future outcomes. Mathematically, ARIMA models the time series data as a linear combination of past observations (AR), past forecast errors (MA), and the differenced data to achieve stationarity (I). The model is represented as ARIMA(p, d, q), where p, d, and q are the orders of the autoregressive, differencing, and moving average components respectively. The AR and MA parts capture the autocorrelations in the data, while the differencing ensures that non-stationary data is transformed into a stationary process, thus making it easier to model. The core intuition behind ARIMA is that the future values of a series can be predicted from the weighted sum of its past values and the forecast errors, assuming that the underlying processes governing the data are linear and temporal.

PCR, on the other hand, is a regression technique that first reduces the dimensionality of the dataset by applying Principal Component Analysis (PCA). PCA transforms the original correlated predictor variables into a new set of uncorrelated variables, known as principal components, which capture the most significant variance in the data. These principal components are then used in the regression model to predict the target variable, in this case, AQI. The intuition behind PCR is that by projecting the data into a lower-dimensional space, we eliminate noise and multicollinearity, which can lead to overfitting in traditional regression models. Mathematically, the process involves first calculating the eigenvectors (principal components) of the covariance matrix of the input data and then projecting the data onto the subspace spanned by the top eigenvectors. The regression model is then fit using these components, ensuring that the relationships between AQI and the predictor variables are captured more efficiently.

The integrated approach of combining ARIMA and PCR leverages the strengths of both methods. ARIMA handles the temporal aspect of the AQI data, capturing the autocorrelation and seasonality inherent in air quality measurements. PCR, in turn, addresses the multivariate nature of AQI prediction by incorporating external factors like meteorological data, traffic density, and industrial emissions. The integration typically involves using ARIMA to model the temporal patterns in the AQI data, and then using PCR to incorporate additional predictors. A practical implementation might involve using ARIMA to forecast the AQI for future time steps, and then using the forecasted values as features in a PCR model, where the principal components of the predictor variables are regressed to refine the AQI prediction. This combined approach is mathematically beneficial as it allows the model to capture both the temporal dependencies in AQI data and the complex relationships between AQI and multiple influencing factors, leading to more accurate and robust predictions. However, the method also introduces computational complexity as both models need to be calibrated and tuned, requiring careful preprocessing and validation to ensure that the integration enhances the prediction accuracy without overfitting.

Methodology

1. Data Acquisition
   In this study, daily air quality data of PM10, PM2.5, NO2, Ozone, CO, SO2 and NH3 over a period of 2010–2023 at multiple stations, obtained from Central Pollution Control Board (CPCB), Delhi has been used.

2. Data Loading
   Multiple .csv files representing data for stations across Delhi was first concatenated to form a single datafile containing the mean values to represent the entire region.

3. Data Cleaning and Preprocessing
   Redundant and insignificant attributes were removed and missing values were handled using fill methods, median values, and interpolation before proceeding with analysis. Interpolation estimates

values within a known data range, while extrapolation predicts values outside it; time series forecasting typically uses extrapolation to predict future data points based on historical trends. Hourly data was then resampled to represent monthly figures based on mean average.

Sharma et al. (2001, 2002, 2003) developed an AQI scale for IIT-Kanpur and India using the Maximum Operator Approach, which determines AQI based on the highest sub-index among all pollutants. Sub-index function represents the relationship between pollutant concentration Xi and corresponding sub-index Ii. It is an attempt to reflect environmental consequences as the concentration of specific pollutant changes. It may take a variety of forms such as linear, non-linear and segmented linear. Typically, the I-X relationship is represented as $I = \alpha X + \beta$ where, α =slope of the line, β = intercept at X=0. The general equation for the sub-index (I_i) for a given pollutant concentration (C_p); as based on 'linear segmented principle' is calculated as:

$$I_i=[\{(I_{HI} - I_{LO})/(B_{HI} - B_{LO})\} * (C_p - B_{LO})]+ I_{LO}$$

where,
B_{HI}= Breakpoint concentration greater or equal to given concentration
B_{LO}= Breakpoint concentration smaller or equal to given concentration
I_{HI} =AQI value corresponding to B_{HI}
I_{LO} = AQI value corresponding to B_{LO}
C_p = Pollutant concentration

(i) Breakpoints for AQI Scale (0-500) (units: $\mu g/m^3$ unless mentioned otherwise)

Table 3.11 Breakpoints for AQI Scale 0-500 (units: µg/m^3 unless mentioned otherwise)

AQI Category (Range)	PM_{10} 24-hr	$PM_{2.5}$ 24-hr	NO_2 24-hr	O_3 8-hr	CO 8-hr (mg/m^3)	SO_2 24-hr	NH_3 24-hr	Pb 24-hr
Good (0-50)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200	0-0.5
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400	0.6 –1.0
Moderate (101-200)	101-250	61-90	81-180	101-168	2.1- 10	81-380	401-800	1.1-2.0
Poor (201-300)	251-350	91-120	181-280	169-208	10.1-17	381-800	801-1200	2.1-3.0
Very poor (301-400)	351-430	121-250	281-400	209-748*	17.1-34	801-1600	1201-1800	3.1-3.5
Severe (401-500)	430 +	250+	400+	748+*	34+	1600+	1800+	3.5+

**One hourly monitoring (for mathematical calculation only)*

CPCB considers reviewing the AQI breakpoints every three years after accounting for research findings on air pollution exposure and health effects.

(iii) Health Statements for AQI Categories

AQI	Associated Health Impacts
Good (0–50)	Minimal Impact
Satisfactory (51–100)	May cause minor breathing discomfort to sensitive people
Moderate (101–200)	May cause breathing discomfort to the people with lung disease such as asthma and discomfort to people with heart disease, children and older adults
Poor (201–300)	May cause breathing discomfort to people on prolonged exposure and discomfort to people with heart disease with short exposure
Very Poor (301–400)	May cause respiratory illness to the people on prolonged exposure. Effect may be more pronounced in people with lung and heart diseases
Severe (401-500)	May cause respiratory effects even on healthy people and serious health impacts on people with lung/heart diseases. The health impacts may be experienced even during light physical activity

(iv) Interpretation of AQI of Delhi for a sample of months in 2010

Date & Time	PM2.5	PM10	NO_2	NH_3	SO_2	CO	Ozone	PM2.5 SI	PM10 SI	NO_2 SI	NH_3 SI	SO_2 SI	CO SI	Ozone SI	AQI
2010-01-31	94.21	213.8	43.99	36.44	8.491	3.514	20.51	211.9	175.9	54.76	9.11	10.61	118.7	20.52	211.9
2010-02-28	97.86	213.8	50.58	21.07	9.562	3.580	22.07	224.4	175.9	63.05	5.27	11.95	119.5	22.08	224.4
2010-03-31	103.9	144.3	35.84	15.40	17.06	2.356	33.64	245.3	129.8	44.81	3.85	21.32	104.2	33.64	245.3
2010-04-30	76.53	55.58	42.15	11.61	16.91	3.505	45.05	154.0	55.59	52.45	2.90	21.14	118.6	45.06	154.0

4. Model Fitting

- Time Series Data

A sequence is an ordered list of events which may be categorised into three groups: time-series data, symbolic sequence data, and biological sequences. A time-series data set consists of sequences of numeric values obtained over repeated measurements of time (typically equal intervals of time such as minutely, hourly, daily etc.). Typically, each time series describes the evolution of an object as a function of time at a given data collection station.

Trend analysis, often used for time-series forecasting, builds an integrated model using the following four major components or movements to characters time-series data:

1. Trend or long term movements: These indicate the general direction in which a time-series graph is moving over time, for example, using weighted moving average and the least squares methods to find trend curves.
2. Cyclic movements: These are the long term oscillations about a trend line or curve.
3. Seasonal variations: These are nearly identical patterns that a time-series appears follow during corresponding seasons of successive years. For effective trend analysis, the data often needs to de seasonalised based on a seasonal index computed by autocorrelation.
4. Random movements: These characterise sporadic changes due to chance events.

The simplest model for a time series is one in which there is no trend or seasonal component and in which the observations are simply independent and identically distributed (iid) random variables with zero mean. We refer to such a sequence of random variables $X_1, X_2, \ldots$ as iid noise.

- White Noise is defined as a sequence of independent, identically distributed random variables with a constant mean and variance. It is typically acquired by a time series with mean=0, correlation between lags = 0, and constant volatility (standard deviation) over time. White noise is considered *unpredictable* because it is entirely random, with no discernible pattern or correlation between its values over time (meaning that knowing past values gives no information about future values). White noise has equal intensity at all frequencies within a given range, resulting in a "flat" power spectral density. This uniform distribution across frequencies means there's no dominant frequency or pattern that could be extrapolated to predict future behavior. Furthermore, each point in white noise is independent of all others.
  yt = signal + noise is a common representation, where yt is the observed data at time t, signal represents the underlying pattern in the data, which may include trends, seasonality, and cyclic components — essentially, the part of the data that can be systematically explained and potentially predicted. Noise is the random component, often assumed to be white noise if it's truly unpredictable, representing the part of the data that cannot be explained by any model. If the residuals (yt-signal) resemble white noise — that is, they are random, uncorrelated, and have zero mean — it indicates that your model has captured all the predictable structure in the data and has attained optimal performance.
  Every IID $(0,\sigma^2)$ sequence is WN $(0,\sigma^2)$ but not conversely. If the time-series data is white noise it implies it is stationary but not conversely.

- Stationarity assumes that any statistical property of the medium (mean, variance, covariance, or higher-order moments, defined by the “ensemble averaging” concept) is stationary in space (i.e., does not vary with a translation): it will be the same at any point of the medium.

  Let $\{X_t\}$ be a time series with $E(X_t^2) < \infty$. The mean function of $\{X_t\}$ is $\mu_X(t) = E(X_t)$. The covariance function of $\{X_t\}$ is $\gamma_X(r,s) = Cov(X_r,X_s)E[(X_r - \mu_X(r))(X_s - \mu_X(s))]$ for all integers r and s.
  $\{X_t\}$ is (weakly) stationary if
  (i) $\mu_X(t)$ is independent of t, and (ii) $\gamma_X(t+h,t)$ is independent of t for each h. The Augmented Dickey Fuller (ADF) test is run to check for data stationarity.

  (A) Dickey Fuller Test

  This test assumes that our time series data is AR1 i.e.
  $y_t = \mu + \phi_1 y_{t-1} + \varepsilon_t$
  $H_o: \phi_1 = 1$

$H_1: \phi_1 < 1$

$y_t - y_{t-1} = \mu + (\phi_1 - 1)\, y_{t-1} + \varepsilon_t$
$\Delta y_t = \mu + \delta\, y_{t-1} + \varepsilon_t$
$H_o: \delta = 0$
$H_1: \delta < 0$

$t_{\hat{\delta}} = \hat{\delta} / \text{se}(\hat{\delta})$

Comparing with the Dickey Fuller Distribution, if
$t_{\hat{\delta}} < DF_{CRITICAL}$: Reject H_0 (data is stationary)
$t_{\hat{\delta}} > DF_{CRITICAL}$: Do not reject H_0

(B) Augmented Dickey Fuller Test

$y_t = \mu + \sum_{i=1}^{p} \phi_1 y_{t-i} + \varepsilon_t$
$\Delta y_t = \mu + \delta\, y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \varepsilon_t$
$H_o: \delta = 0$
$H_1: \delta < 0$
$t_{\hat{\beta}_i} = \hat{\beta}_i / \text{se}(\hat{\beta}_i)$

The data was initially non-stationary; hence, first-order differencing was applied to achieve stationarity. Depending on the nature of the data, transformations such as logarithmic scaling, Box-Cox transformation, or detrending may also be considered.

- Autoregressive Integrated Moving Average (ARIMA)

ARIMA (AutoRegressive Integrated Moving Average) is a statistical model used for time series forecasting. It combines three components to make predictions:

AR (AutoRegressive): This component models the relationship between an observation and a specified number of lagged observations (previous time points). It uses the dependency between current and past values to predict future values. Let $\{X_i\}$ be a stationary time series. The auto covariance function (ACVF) of $\{X_t\}$ at lag h is $\gamma_x(h) = \text{Cov}(X_{t+h}, X_t)$.
The autocorrelation function (ACF) of $\{X_i\}$ at lag h is $\rho x(h) = \gamma_x(h)/\gamma_x(0) = \text{Cov}(X_{t+h}, X_t)$ACF measures the correlation between a time series and its past values (lags) at various time intervals, indicating how each lag is related to the series. PACF (Partial Autocorrelation Function) measures the direct correlation between a time series and a specific lag, controlling for the influence of any shorter lags in between.

I (Integrated): This step involves differencing the series to make it stationary, meaning its statistical properties (like mean and variance) do not change over time. Differencing helps to remove trends and seasonality from the data.

MA (Moving Average): This component models the relationship between the observation and the residual errors from a moving average model applied to lagged observations. In this model, Y_t depends on the current residual error ϵ_t as well as the residual errors of previous time points.

The model is denoted as ARIMA(p, d, q), where:

- p is the number of lag observations (AR),

- d is the degree of differencing (I),
- q is the size of the moving average window (MA).

the values of p, d, and q in an ARIMA model are often selected based on model evaluation criteria like the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC).

AIC is a statistical measure that helps in comparing different models. It takes into account the goodness of fit (how well the model explains the data) and the complexity of the model (penalizing for having too many parameters).

$$AIC = 2k - 2\ln(L)$$

where k: Number of parameters in the model
L: Maximum likelihood of the model

Lower AIC values indicate a better model (one that fits the data well with fewer parameters). When comparing multiple ARIMA models with different values of p, d, and q, the one with the lowest AIC is generally preferred.

5. Forecasting
Once the model is fit onto the training data, predictions were made on the test data. The ratio to training set to test set was set at 80:20. The following performance measures were used to test model accuracy: Mean Absolute Error, Mean Squared Error, Mean Absolute Percentage Error, Mean Absolute Scaled Error, Ljung-Box test, and Normal Test.

Experimental Results and Discussion

Further, AQI for future months was predicted.

(i) Predicted AQI

Year/Month	Predicted AQI
2023-04-30	225.469007
2023-05-31	213.945976
2023-06-30	258.736044
2023-07-31	232.051845
2023-08-31	268.234468
2023-09-30	237.384479
2023-10-31	270.871769
2023-11-30	239.028481
2023-12-31	271.529189
2024-01-31	239.606617
2024-02-29	271.615950

2024-03-31	239.876098
2024-04-30	271.538892
2024-05-31	240.055541
2024-06-30	271.415463
2024-07-31	240.208071
2024-08-31	271.279562
2024-09-30	240.351919
2024-10-31	271.140967
2024-11-30	240.492358
2024-12-31	271.002493
2025-01-31	240.630918
2025-02-28	270.864945
2025-03-31	240.768048
2025-04-30	270.728546

(ii) Performance Measures

Performance Measure	**Value**
Absolute Error (MAE)	99.98
Mean Squared Error (MSE)	13331.33
Mean Absolute Percentage Error (MAPE)	48.57%
Mean Absolute Scaled Error (MASE)	1.02
Normality Test p-value	0.2384
Autocorrelation of Residuals p-value	> 0.05 (Uncorrelated)

Absolute Error (MAE): 99.98, indicating that, on average, the model's predictions are off by approximately 100 units.

Mean Squared Error (MSE): 13331.33, which reflects the average squared difference between the predicted and actual values. This suggests that, although the model captures the overall trend, there is some significant deviation.

Mean Absolute Percentage Error (MAPE): 48.57%, meaning that the model's predictions are, on average, about 48.57% away from the actual values, highlighting that the model's predictions may be relatively inaccurate in certain cases.

Mean Absolute Scaled Error (MASE): 1.02, which is close to 1, suggesting that the model performs comparably to a naive forecast model, where a value greater than 1 indicates the model is less accurate than a naive forecast.

Normality Test p-value: Based on the Jarque-Bera test a value of 0.2384 indicates that the residuals from the model are likely normally distributed, which is a good sign for the model's assumptions.

Autocorrelation of Residuals: The p-value for the residuals autocorrelation test is greater than 0.05, suggesting that the residuals are uncorrelated, meaning the model has successfully captured the underlying structure of the data without significant patterns remaining in the residuals.

Conclusion and Future Scope

The future development of this project aims to address key areas that can enhance the accuracy, reliability, and robustness of the ARIMA-based model for Air Quality Index (AQI) prediction. The following objectives outline the areas of improvement and expansion:

1. Addressing Heteroscedasticity through GARCH-ARIMA Integration: One of the challenges faced in time series forecasting, particularly with the ARIMA model, is the potential presence of heteroscedasticity in the residuals. Heteroscedasticity occurs when the variance of errors is not constant over time, which can lead to suboptimal forecasts and unreliable predictions. To address this, the integration of Generalized Autoregressive Conditional Heteroscedasticity (GARCH) models with ARIMA will be explored. GARCH models are particularly effective in modeling changing variance over time, and by combining them with ARIMA, we aim to improve the model's performance in forecasting AQI by adjusting for periods of high or low volatility. This integration would refine the model by accounting for the dynamic nature of variance in the residuals, leading to more accurate and reliable predictions.

2. Understanding and Addressing Cyclicity in ARIMA Predictions: Cyclical patterns or seasonal variations in the AQI data could lead to periodic fluctuations that the ARIMA model might fail to capture entirely. These cyclical trends could result in errors when forecasting AQI values during certain times of the year or specific conditions. Future work will focus on identifying and understanding the cyclical nature of the AQI time series data. Advanced techniques, such as seasonal decomposition or the use of SARIMA (Seasonal ARIMA), could be explored to separate seasonal components from the data and address cyclicity. By doing so, we aim to eliminate cyclical biases in the model’s predictions, ensuring more precise forecasting, especially during times of high variability in AQI levels.

3. Incorporating Exogenous Variables (Wind, Temperature, Gusts, etc.): Currently, the model focuses primarily on the historical AQI data for prediction. However, external factors such as wind speed, temperature, humidity, and gusts play a significant role in influencing air quality and should be incorporated into the forecasting model. By integrating these exogenous variables (often referred to as external regressors) into the ARIMA framework, we can enhance the model’s ability to account for these environmental factors that directly impact AQI levels. For instance, high wind speeds might disperse pollutants, reducing AQI levels, while high temperatures could exacerbate pollution levels. The inclusion of these factors will provide a more comprehensive understanding of the drivers behind AQI fluctuations, enabling more accurate predictions. This could be achieved by using ARIMAX (ARIMA with Exogenous Variables) models, which would allow the model to factor in these additional inputs and potentially improve forecasting accuracy by addressing the interactions between AQI and external factors.

References

Central Pollution Control Board. (n.d.). *National air quality index*. Retrieved November 27, 2024, from https://cpcb.nic.in/displaypdf.php?id=bmF0aW9uYWwtYWlyLXF1YWxpdHktaW5kZXgvRklOQUwtUkVQT1JUX0FRSV8ucGRm

Central Pollution Control Board. (n.d.). *AQI repository*. Retrieved November 27, 2024, from https://airquality.cpcb.gov.in/ccr/#/caaqm-dashboard-all/caaqm-landing/aqi-repository

Borsuk, M. E., Stow, C. A., & Sutherland, R. A. (2011). *Spatio-temporal prediction of air quality using land-use and demographic data*. *Science of the Total Environment*, *409*(6), 1170-1183. https://doi.org/10.1016/j.scitotenv.2010.11.059

Brockwell, P. J., & Davis, R. A. (2008). Introduction to time series and forecasting (2nd ed.). Springer.